Exhibit A

The Home Video System of the Future

Rainer Lienhart

Intel Research Labs
Santa Clara, CA 95052
{Rainer.Lienhart, Bob.Davies}@intel.com

ABSTRACT

1 Introduction/Motivation

Three basic motivation for creating, maintaining and browsing a home video archive are identified:

1. Archiving: The recorder of the material and/or his/her close friends want to perserve their memories without any loss. Thus the raw unedited video material should be stored.
2. Browsing: Since the content of any archive might be used some time later, e.g. for the preparation of an aniversary talk about, video browsing should be supported
3. Entertainment/Communication: Often, home video material show the clips from the peoples' vacation. Many times the people do not have the time to edit their raw video material, thus it is to long and boring. Thus, a automatic abstracting procedure is needed that generate video abstracts of the hightlight automatically. To do that the video and audio track must be analysed. Since often only very little semantically important information can be extracted form those tracks, easy video anontation during recording is allowed
4. Edited Presentation: The video material is arranged in an adequate manner

2 System Overview

2.1 Modern Video Aquisition

If you look at the current process of home video aquisition, the people just record the desired actions/events with their consumer camera without adding any highlevel or context information. They maybe luckly and own a good camera so they get the time and date of recoding, but that is basically all.

By slightly changing the way people capture home video material, a lot of highlevel and context information can be added effortless during video aquisition enabling so far impossible, but desired video processing possibilities.

For instance, if you are on a sightseeing tour you could speak out the name of the place and object you are recording loud. Two issues have to be considered.

1. In most cases, those annotations are unwanted noise in your audio during video playback. Thus, we must think about ways to get rid of those annotations during playback.
2. In general, public places have a lot of background audio making it difficult for voice recognition tool to transscribe the annotation to ASCII.

To solve that problem, we propose to use two microphones during video aquisition (see also Figure xx):

- A unidirectional dynamic headworn microphone designed for computer speech input applications (henceforth called voice microphone) capturing only the voice of the speaker and cutting out everything else and

- a general unidirectional microphone to capture scene sound (henceforth called scene microphone) capturing the voice of the speaker and anything else.

In principle, the output of both mics should be stored in separate audio channels. However, this is currently supported only by a few expensive DV cameras. Thus, to allow our setup to work with any consumer camera, the output of the voice microphone is stored as the left audio channel, while the output of the scene microphone is stored as the right audio channel (see Figure xx).

Due to the specific features of the voice microphone the right channel can be used directly as input to a commercial speech recognition tool such Dragon Naturally speaking and ViaVoice. We therefore turn the focus not to the question of how to remove the annotations from scene sound.

2.1.1 Annotation Removal

In order to remove the annotations from the scene sound, the must find a way to predict from the voice sound how the annoation would sound on the scene sound. Let assume that the cameraman talks for ten seconds in an empty room

Let $v(t)$ be the signal recorded by the voice microphone and $s(t)$ the one recorded by the scene microphone. What we know want to extract from the training samples is the filter f(t) that minimizes the expression

$$(v(t) * f(t) - s(t))^2 \rightarrow min$$

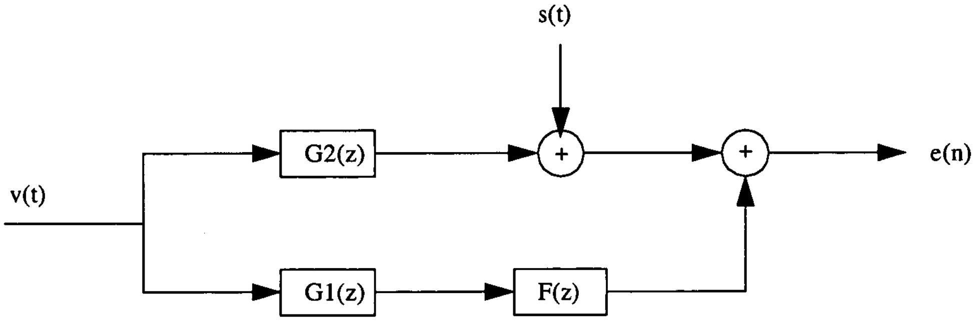


Figure 1: Adaptive filtering approach for echo cancellation

2.1.2 Annotation Language

2.1.3 Recording Mode

Two different audio channels are available: Scene sound and annotations/Camera commands (like stop, record, delete last recording). The annotation sound should be substracted from the sceen sound. The video records the scene.

2.1.4 Learn Mode

Kontrollsprache genau entwickeln

Learn New face

Firstly, the background is recorded without any face; Then, the subject rotates 360 or +-90 degree in front of the camera.

- Learn object

2.2 Storage

To views: * raw video

* set of edits (like views on the video)

2.2.1 Extraction of Time and Date

FROM ANALOG VIDEO: Extract

FROM DIGITAL VIDEO: In the case of digital video cameras the time and date can be retrieved via the IEEE1394 interface (see [19]).

2.2.2 Shot Clustering Based on Time and Date

From our experiences with time-constrained clustering algorithms [][] we know that temporal distance plays an important role in shot clustering. All current shot clustering algorithms so far, however, measure the temporal distance by the time code difference of respective frames or shots. Thus, the notion of temporal distance refer to the distance of the playback time. While this might be appropriate for broadcasts such as feature films, commercials, sitcom etc., where the time and date of recording in most cases is of useless semantic, the difference in the time and date of recording - if avaiable - seems to be the primary choice for home video. And with the advent of DV cameras, this information can easily be retrieved for each frame.

Two contiguous shots $S_i = [t_b^i, t_e^i]$ and $S_{i+1} = [t_b^{i+1}, t_e^{i+1}]$ recorded closely in time are very likely to show the same event, locale or action. A common measure for temporal distance between two shots $S_i = [t_b^i, t_e^i]$ and $S_j = [t_b^j, t_e^j]$ is the time elapsed between them, i.e.

$$\Delta t(S_i, S_j) = \begin{cases} 0 & if\ [t_b^i, t_e^i] \cap [t_b^j, t_e^j] \neq \varnothing \\ t_b^j - t_e^i\ if\ t_b^j > t_e^i \\ t_b^i - t_e^j\ if\ t_b^i > t_e^j \end{cases} \quad . \tag{1.1}$$

In general, two shots are put into the same cluster if they are not apart more than a certain time.

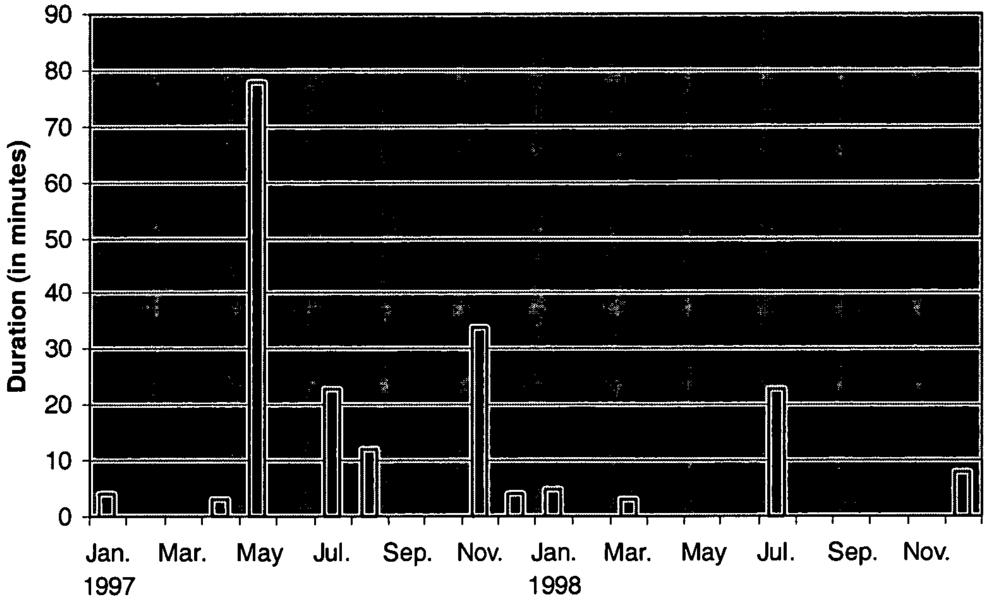
This leads us to the following hierarchical clustering algorithm:

1. Group all shots which are less apart than 5 minutes (henceforth *level 1 clusters*).
   Since you cannot go or get very far within 5 minutes, the contents of the shots within level 1 cluster should belong to the same locale or should show related actions. Note, however, that the same locale or related actions may be distributed over several level 1 clusters.
2. Group all shots which are less apart than 1 hour (henceforth *level 2 clusters*).
   The choice of the threshold is motivated by grouping together events such as birthdays parties, weddings, or playing outside in the snow. Thus, level 2 clusters represents locales and actions in a wider sense.
3. Group all shots which are less apart than 6 hours (henceforth *level 3 clusters* or *day clusters*).
   Assuming that people usually sleep at least 6 hours per day, level 3 clusters should group shots into "individual days". A day in this context denotes the time between consecutive night rests. It does not imply a single calender date.
4. Group all shots which are less apart than 5 days (henceforth *level 4 clusters* or *vacation clusters*).
   The motivation for the choice of the threshold is twofolded. Firstly, you can assume that during vacations the duration between to contiguous recordings should be shorter than 5 days. Secondly, is allows to group weekend and long weekends while avoiding to group consecutive weekends if they are not bridged by

recordings during the work days.

On our home video archive comprising xx hours of MPEG home video over a periods of yy year, the clusters coincident quite well with their intended semantics. Note, that in general one cannot expect that level 2 to 4 clusters we have defined may be determined by their audio-visual similarity since their are usually to far apart to have any audio-visual similarity.

5. Show small bars on a time line (gives visual clusters, the bar hight represents the length in that period; or better color images (d.h. verschieden Haeufigkeiten bekommen verschiedene Farben (wie Wetterkarten) und jedes Streifen repraesentiert ein Jahr; weiss = 0; rot=am haeufigsten)) => Eigentlich gehoert dies zu browse mode; Kann auch dazu genommen werden, um bei Nachrichten die Wichtigkeit einer Nachricht zu beurteilen. Z.B. bei Bill Clinton. Wie viel Video wurde ueber ihn zu welcher Zeit gesendet. Z.B. kann man dann unterscheiden zwischen Zwischenergebnis und endgueltigen Ergebnissen in einer Sache. Wenn sehr viel gesendet wird, dann weiss man eventuell noch nicht genug darueber, und man sollte am Ende des Peaks einsteigen, um sich zu informieren.



2.3 Playback

2.3.1 Video Abstracts

Commonly, the term "video abstract" denotes a sequence of moving images which presents the content of a much longer video in such a way that the respective target group is rapidly provided with concise information about the content. One key point of abstracts is that they are much shorter than the original.

In previous work we was devoted to abstracting feature films automatically. This time, however, we concentrate on home video material which is inherently different from all "commercial" video material prompting the demand for new abstracting principles and algorithms.

The differences between home videos and feature films are:

- Feature films tell stories, have a plot and follow certain design rules. Most of them present several side stories around the main story. Consequently, some shots/scenes are more important than others, suggesting immediately that only the important shots should be part of a video abstract. Contrary to that, home videos document the life of the people. There is no artificial story, plot or structure in the video material leading us to the hypothesis that all shots are more or less equally important.
- Home videos are inherently time-stamped data. Time and date of recording are very important attributes of the video material. Current camcorders store them on the tape along with the video. This quality does definitely not apply to feature films. For feature film the sense of time and date is constructed within the video by means of the story. The time and date of recording is only misleading.
- The duration of feature films is restricted lasting usually between 1.5 and 2.5 hours. On the other hand, since home videos document the people's life and are becoming more and more popular, the total duration of recorded material easily sum up to hundreds of hours within a few years.
- Feature films are edited by professionals. Previous work made use of that knowledge by always selecting complete shots as clips for a video abstract. Home video, however, is unedited, raw footage raising the demand to find some reasonable in and out point in order to shorten the individual shots.
- Home video is not shot for a general audience, but for friends, guests and oneself.

Taking these observations and the fact into account that the target group of home video are the family members and friends, led us to the following rules for good-quality abstracts:

1. *Balanced coverage.* The video abstract should be composed of clips distributed over the souce video set.
2. *Shorten shots.* Commonly the duration of the raw, unedited shots is too long and too long-winded for an abstract. Moreover, its unedited present in a video abstracts endangers the balanced coverage. Thus, long shot must be cut down to its most interesting part.
3. *Select randomly.* As already mentioned, due to the nature of home video material, shots are more or less equally important. Moreover, individual abstracts of the same source videos should vary to order to be still interesing to the owner of the archive. Therefore, random slip selection represents an important part of the video abstracting procedure.
4. *Vary abstracting principles.* If different event are covered by the video abstract each event should be covered by different randomly selected abstracting principles

Our Approach:

The user specifies the time period the abstract should cover as well as its duration. Dependent on the ratio of duration of the source video material and abstract different strategies are applied. In order to explain whose strategies we have to introduce our abstarcting model.

* different events

* reduce play time by cut down + transitions.

The abstracting algorithm we have developed can be subdivided into three consecutive steps (see Figure 2). In step 1, video segmentation and analysis, the input video is segmented into its shots and scenes. At the same time frame sequences with special events, such as text appearing in the title sequence, close-up shots of the main actors, explosions, gun fires, etc. are determined. In the second step, clip selection, those video clips are selected which will become part of the abstract. The third step, the clip assembly, assembles the clips into their final form and produces the presentation layout. This implies determining the order of the video clips, the type of transitions between them, and other editing decisions.

Select Clip:

Target Group

The audience of home video material can roughly be divided into two groups:

- Friends, Guest and other third party people and: delete similar video material to shorten video (important ; is a editing rule!)-> dynamic
- Recorder or its close friends: delete only very little information -> dynamic

Need automatic editing rules to make content more interesting; Annotation track mit einblenden als Text?

For instance, if a person in approaching and it takes a long time -> speed up approach by serveral dissolves with time jumps!

Pans: Stetch pans together and let videoplayer move over the image, i.e. you see a static panorama there the object in the current video frame starts to move.

- Indexing -> paper, static
- Friends -> paper, static

2.3.2 Classify shots

- blured
- unintentionally recorded
- Peolpe are present:
  - Size: at least 20x20; Recognize view + Person (Note: Research seems to propose that face recognition, especially frontal faces, are done in a dedicated location of the brain).
  - Size: smaller than 20x20, whole Person visible => train texture pattern
  - Size: smaller than 20x20, crowd visible => train texture pattern

Annotation

Use annotations as crystal points for shot clustering / story units and start to grow from those crystal points. First of all group all shots with no or little time gap (less than 5 minutes). Second level grouping

Linearly:

-
- Same video/audio setting
- Same day

Non-Linearly:

- Same person
- Similar content

2.3.3 Editing Support

Scene Grid

- Edge register frame with in shots, in order to remove hand jitter. Allow to specifiy sequence of bounding boxes. The boxes inbetween are interpolated.

3 Experimental Results

4 Conclusion and Future Research Direction

References

[1] *Adobe Premiere 4.0 Handbuch*, Adobe Systems, San Jose, CA, USA, 1995

[2] P. Aigrain und P. Joly. The Automatic Real-Time Analysis of Film Editing and Transition Effects and Its Applications. *Computer and Graphics*. Vol. 18, No. 1, pp. 93-103, 1994.

[3] J. S. Boreczky and L. A. Rowe. Comparison of Video Shot Boundary Detection Techniques. In *Storage and Retrieval for Still Image and Video Databases IV*, Proc. SPIE 2664, pp. 170-179, Jan. 1996.

[4] J. Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 6, pp. 34-43, Nov. 1986.

[5] A. Dailianas, R. B. Allen, P. England: Comparison of Automatic Video Segmentation Algorithms. In *Integration Issues in Large Commercial Media Delivery Systems*, Proc. SPIE 2615, pp. 2-16, Oct. 1995.

[6] U. Gargi, R. Kasturi, and S. Antani. Performance Charactization and Comparison of Video Indexing Algorithms. Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 559-565, June 1998.

[7] A. Hampapur, R. C. Jain, and T. Weymouth. Production Model Based Digital Video Segmentation. *Multimedia Tools and Applications*, Vol.1, No. 1, pp. 9-46, Mar.1995.

[8] R. Lienhart. Methods ..******....... Ph.D. thesis, University of Mannheim, Juli 1998.

[9] R. Lienhart, C. Kuhmünch, and W. Effelsberg. On the Detection and Recognition of Television Commercials. In *Proceedings of the International Conference on Multimedia Computing and Systems*, Ottawa, Ontario, Canada, pp. 509-516, June 1997.

[10] A. Nagasaka and Y. Tanaka. Automatic Video Indexing and Full-Motion Search For Object Appearances. In Proc. *Second Working Conf. on Visual Databases Systems*, pp. 113-127, Sept. 1991.

[11] O. Otsuji and Y. Tonomura. Projection Detecting Filter for Video Cut Detection. Proc. *First ACM International Conference on Multimedia*, pp. 251-257, 1993.

[12] K. Otsuji, Y. Tonomura, and Y. Ohba. Video Browsing Using Brightness Data. In *Visual Communication and Image Processing*, Vol. SPIE 1606, pp. 980-989, 1991.

[13] B.-L. Yeo and B. Liu. Rapid Scene Analysis on Compressed Video. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 5, No. 6, December 1995.

[14] J. Yu, G. Bozdagi, and S. Harrington. In *Proc. International Conference on Image Processing*, Vol. 2, pp. 498-501, 1997

[15] R. Zabih, J. Miller, and K. Mai. A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. *Proc. ACM Multimedia 95*, San Francisco, CA, pp. 189-200, Nov. 1995.

[16] H. J. Zhang, A. Kankanhalli, and S. Smoliar. Automatic Partitioning of Full-Motion Video. *Multimedia Systems*, Vol. 1, No. 1, pp.10-28, 1993.

[17]

[18]

[19] AV/C Digital Interface Command Set VCR Subunit Specification. 1394 Trade Association Steering Committee. Version 2.0.1, January 5, 1998

[20]